Degrees of Freedom of the Generalized Least Squares Error

Bill Menke, October 16, 2014

1. The generalized least squares error is $E_g = E_d + E_h$ with data error $E_d=||(\mathbf{d}\text{-}\mathbf{Gm})/\sigma_d||^2$ and prior information error $E_h=||(\mathbf{h}\text{-}\mathbf{Hm})/\sigma_h||^2$, for N data **d** and K pieces of prior information **h**.

2. Here we examine a problem with N=21 and K=5, based on fitting a degree 6 polynomial to 21 data.

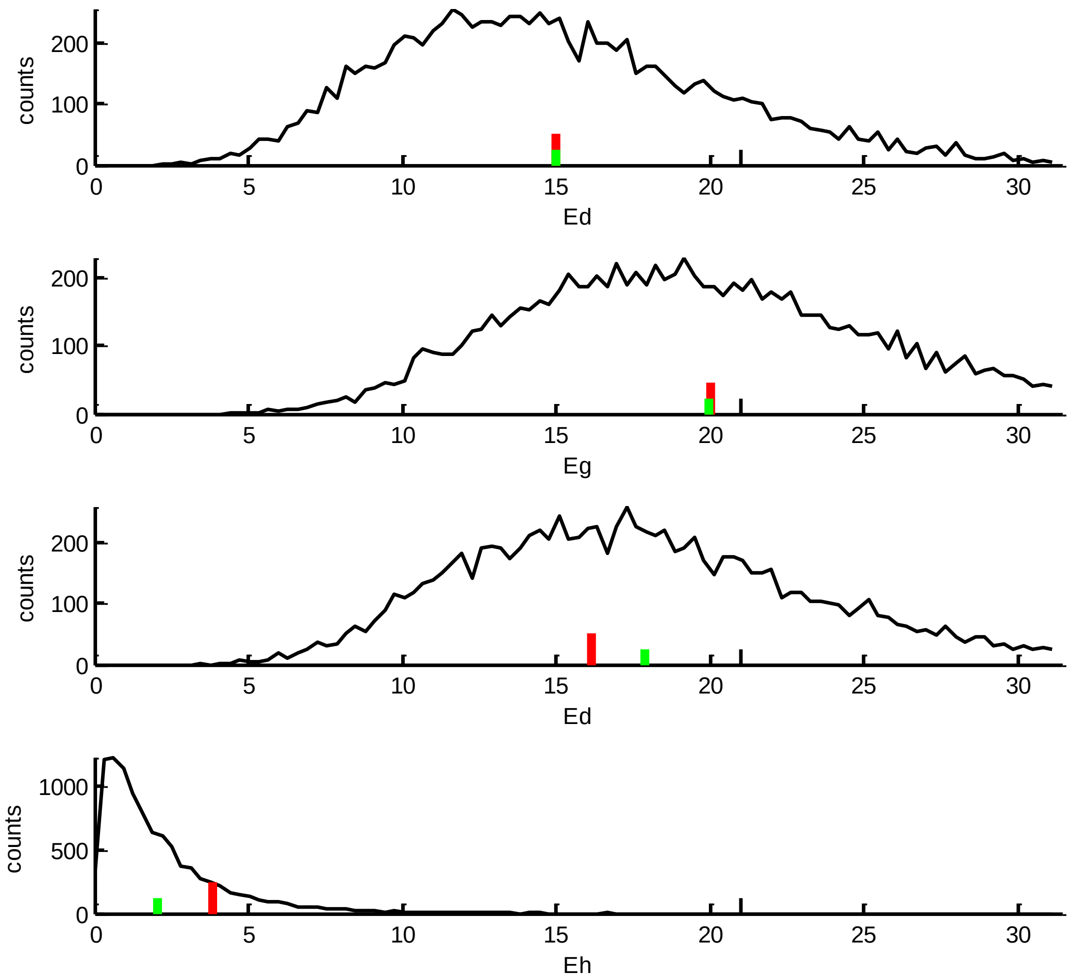
3. For ordinary least squares, the degrees of freedom are ν=N-M=21-6=15. The Monte-Carlo simulation in the top panel of the figure shows a histogram of E_d, which follows a chi-squared distribution with ν degrees of freedom, which has a mean value of ν. Note that the predicted (red bar) and estimated mean (green bar) degree closely, and are clearly to the left of N=21 (black bar). A chi-squared test on the significance of E_d being different from the expected value would clearly work.

4. We add K=5 constraints to the problem, representing knowledge of the difference between adjacent model parameters.

5. For generalized least squares, the degrees of freedom are ν=N+K-M=21+5-6=20. The Monte-Carlo simulation in the second-from-top panel shows a histogram of E_g, which follows a chi-squared distribution with ν degrees of freedom, which has a mean value of ν. Note that the predicted (red bar) and estimated mean (green bar) degree closely.

6. We also plot the histogram of E_d and of E_h (third and fourth panels from top). Consider, first, the histogram of E_d. Note that it is shifted to the right compared to the one in the top panel. Generalized least squares spreads the loss of M degreesof freedom associated with the model parameters across N+K "observations", a larger number than the N of ordinary least squares, so the number of degrees of freedom moves closer to N=21. The Welch–Satterthwaite equation[1] indicates that the ν degrees of freedom of the N+K observations are to be spread proportionately across the N data and K pieces of prior informtion: $\nu_d=\nu N/(N+K)$ and $\nu_h=\nu K/(N+K)$. As can been seen from panels 3 and 4, the approximation (red bar) does not match the estimate (green bar) very well. A chi-squared test on the significance of E_g would clearly work; whether a corresponding test on E_d or E_h could be believed is dubious.

[1]Welch–Satterthwaite equation , Wikipedia, en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite_equation, 2014.



counts
200
100
0
0
5
10
15
20
25
30
Ed
counts
200
100
0
0
5
10
15
20
25
30
Eg
counts
200
100
0
0
5
10
15
20
25
30
Ed
counts
1000
500
0
0
5
10
15
20
25
30
Eh

```
clear all;

N=21;
L=2;
Dx=1;
x=L*[0:N-1]'/(N-1);

mtrue = [5, 4, 3, 2, 1, 0.5]';
M = length(mtrue);

G = [ones(N,1), x, x.^2, x.^3, x.^4, x.^5];
dtrue = G*mtrue;
sigmad = 0.01;

% PART 1: Ordinary Least Squares

Nr = 10000;
Ed = zeros(Nr,1);
nu = N-M;
Emeantrue = nu;
Estdtrue = sqrt(2*nu);
Nbins = 101;
Edmax = (Emeantrue+3*Estdtrue);
bins = Edmax*[0:Nbins-1]'/(Nbins-1);
plotxmax = Edmax;
```

```
GTGinv = inv(G'*G);
for ir=[1:Nr]
    dobs = dtrue + random('Normal',0,sigmad,N,1);
    mest = GTGinv*(G'*dobs);
    dpre = G*mest;
    ed = (dobs-dpre)/sigmad;
    Ed(ir) = ed'*ed;
end

Emeanest = mean(Ed);
myhist = hist(Ed,bins)';
plotymax = max( myhist(1:end-1) );
figure(1);
clf;
subplot(4,1,1);
set(gca,'LineWidth',2);
hold on;
axis( [0, plotxmax, 0, plotymax] );
plot( bins(1:end-1), myhist(1:end-1), 'k-', 'LineWidth', 2 );
plot( [Emeantrue, Emeantrue]', [0, 0.2*max(myhist)]', 'r-', 'LineWidth', 4 );
plot( [Emeanest, Emeanest]', [0, 0.1*max(myhist)]', 'g-', 'LineWidth', 4 );
plot( [N, N]', [0, 0.1*plotymax]', 'k-', 'LineWidth', 2 );
xlabel( 'Ed' );
ylabel( 'counts' );
```

```
% PART 2: Generalized Least Squares
if( 1 )
    H = [1, -1, 0, 0, 0, 0;
     0, 1, -1, 0, 0, 0;
     0, 0, 1, -1, 0, 0;
     0, 0, 0, 1, -1, 0;
     0, 0, 0, 0, 1, -1];
else
    H = eye(6,6);
end
htrue = H*mtrue;
sigmah = 0.01;
K = length(htrue);
F = [G/sigmad; H/sigmah];
ftrue = [dtrue/sigmad; htrue/sigmah];

nug = (N+K)-M;
Nr = 10000;
Ed = zeros(Nr,1);
Eh = zeros(Nr,1);
Eg = zeros(Nr,1);
Egmeantrue = nug;
Egstdtrue = sqrt(2*nug);
Nbins = 101;
```

```
Egmax = (Egmeantrue+2*Egstdtrue);
bins = Edmax*[0:Nbins-1]'/(Nbins-1);

FTFinv=inv(F'*F);
for ir=[1:Nr]
    dobs = dtrue + random('Normal',0,sigmad,N,1);
    hprior = htrue + random('Normal',0,sigmah,K,1);
    f = [dobs/sigmad; hprior/sigmah];
    mest = FTFinv*(F'*f);
    dpre = G*mest;
    ed = (dobs-dpre)/sigmad;
    Ed(ir) = ed'*ed;
    hpre = H*mest;
    eh = (hprior-hpre)/sigmah;
    Eh(ir) = eh'*eh;
    Eg(ir) = Ed(ir) + Eh(ir);
end

Edmeanest = mean(Ed);
Ehmeanest = mean(Eh);
Egmeanest = mean(Eg);
myEdhist = hist(Ed,bins)';
myEhhist = hist(Eh,bins)';
myEghist = hist(Eg,bins)';
```

```
subplot(4,1,2);
set(gca,'LineWidth',2);
hold on;
plotymax = max( myEghist(1:end-1) );
axis( [0, plotxmax, 0, plotymax] );
plot( bins(1:end-1), myEghist(1:end-1), 'k-', 'LineWidth', 2 );
plot( [Egmeantrue, Egmeantrue]', [0, 0.2*plotymax]', 'r-', 'LineWidth', 4 );
plot( [Egmeanest, Egmeanest]', [0, 0.1*plotymax]', 'g-', 'LineWidth', 4 );
plot( [N, N]', [0, 0.1*plotymax]', 'k-', 'LineWidth', 2 );
xlabel( 'Eg' );
ylabel( 'counts' );

Edmeantrue = nug*(N/(N+K)); % Welch–Satterthwaite approximation
subplot(4,1,3);
set(gca,'LineWidth',2);
hold on;
plotymax = max( myEdhist(1:end-1) );
axis( [0, plotxmax, 0, plotymax] );
plot( bins(1:end-1), myEdhist(1:end-1), 'k-', 'LineWidth', 2 );
plot( [Edmeantrue, Edmeantrue]', [0, 0.2*plotymax]', 'r-', 'LineWidth', 4 );
plot( [Edmeanest, Edmeanest]', [0, 0.1*plotymax]', 'g-', 'LineWidth', 4 );
plot( [N, N]', [0, 0.1*plotymax]', 'k-', 'LineWidth', 2 );
xlabel( 'Ed' );
ylabel( 'counts' );
```

```
Ehmeantrue = nug*(K/(N+K)); % Welch–Satterthwaite approximation
subplot(4,1,4);
set(gca,'LineWidth',2);
hold on;
plotymax = max( myEhhist(1:end-1) );
axis( [0, plotxmax, 0, plotymax] );
plot( bins(1:end-1), myEhhist(1:end-1), 'k-', 'LineWidth', 2 );
plot( [Ehmeantrue, Ehmeantrue]', [0, 0.2*plotymax]', 'r-', 'LineWidth', 4 );
plot( [Ehmeanest, Ehmeanest]', [0, 0.1*plotymax]', 'g-', 'LineWidth', 4 );
plot( [N, N]', [0, 0.1*plotymax]', 'k-', 'LineWidth', 2 );
xlabel( 'Eh' );
ylabel( 'counts' );
```